Identifying False News Through Machine Learning Ensemble Techniques

KANDIMALLA SASANK SAI
STUDENT
CSE
DHANEKULA INSTITUTE OF ENGINEERING AND TECHNOLOGY, VIJAYAWADA, INDIA

Abstract - The World Wide Web's introduction and the quick uptake of social media sites like Facebook and Twitter opened the door to a level of information sharing never seen in human history. Social media platforms are being used by consumers to create and share more information than ever before, some of it false and unrelated to reality. It is difficult to automatically classify a text article as misleading or disinformation. Even a subject-matter expert must consider a variety of factors before determining if an article is true. In this work, we propose to automatically classify news articles using an ensemble machine learning approach. Our research examines many textual characteristics that can be utilized to discern authentic content from counterfeit. We train a variety of machine learning algorithms employing those properties in conjunction with different ensemble approaches, and we assess their performance on four real-world datasets. Our suggested ensemble learner technique outperforms individual learners, as confirmed by experimental evaluation.

Index Terms – false news, machine learning, algorithms

I. INTRODUCTION (HEADING 1)

Both supervised and unsupervised learning techniques have been used to text classification in the present fake news corpus on several occasions [20, 21]. The majority of the research, however, focuses on certain datasets or domains, with the politics domain being the most well-known [10, 19, 21]. As a result, when exposed to articles from different domains, the algorithm that was trained performs less well than it should on the sort of content that it was designed for. The linguistic structure of articles varies across different domains, making it challenging to develop a general algorithm that performs well across all news domains. In this research, we suggest a machine learning ensemble strategy as a solution to the false news identification problem. Our research investigates many textual characteristics that may be utilised to differentiate authentic information from counterfeit. We use such qualities to train a variety of ensemble methods—some of which are not well-examined in the existing literature—using a mix of different machine learning techniques. Due to the learning models' propensity to lower error rates through the use of strategies like bagging and boosting, ensemble learners have shown to be beneficial in a wide range of applications [22]. These methods enable the

efficient and successful training of various machine learning algorithms. We also carried out in-depth tests on four real-world datasets that are accessible to the public. Using the four widely used performance indicators, the findings confirm that our suggested approach performs better than before.

II. ALGORITHMS

To assess the efficacy of false news detection classifiers, we combined our suggested technique with the following learning algorithms.

Regression using Logistic Regression:

Logistic regression (LR) models are used to classify text based on a large feature set and provide a binary output (true/false or real article/fake article), since they offer an easy equation for categorising issues into multiple or binary groups [27]. To achieve the best results for each unique dataset, we did hyperparameter tuning. Several parameters are tried in order to obtain the greatest accuracies from the LR model. The logistic regression hypothesis function has the following mathematical definition [27]: The goal of logistic regression is to minimise the cost function in order to obtain an ideal probability. It does this by using a sigmoid function to convert the output to a probability value. The cost function computation is displayed.

Vector Machine Support:

Another model for binary classification problems is the support vector machine (SVM), which comes with a number of kernel functions [28]. To classify data points, an SVM model's goal is to estimate a hyperplane, also known as a decision border, based on the feature set [29]. The number of features affects the hyperplane's dimension. Since there are several ways for a hyperplane to exist in an N-dimensional space, the goal is to find the plane that has the greatest margin of separation between the data points of two classes.

Layered Perceptron:

An artificial neural network including an input layer, one or more hidden layers, and an output layer is called a multilayer perceptron (MLP). MLP can be as basic as having all three layers, but in our tests, we have optimised the model by adjusting its parameters and layer count to provide the best possible prediction. The code below illustrates how a simple multilayered perceptron model with one hidden layer may be expressed.

KNN, or K-Nearest Neighbors:

KNN is an unsupervised machine learning model that can predict a result on a given set of data without the need for a dependent variable. We give the model enough training data and allow it to determine which specific neighborhood a given data point belongs to. A new data point is allocated to the class with the closest distance if the value of K is 1, which indicates the majority of the votes cast by its neighbors. The KNN model calculates the distance between a new data point and its nearest neighbors. The following mathematical formulas can be used to determine the approximate distance between two places [31].

Group Education:

We proposed to enhance the overall accuracy of classifying an article as true or false by utilizing textual features as feature input in conjunction with existing ensemble methodologies. Because several models are trained using a specific approach to lower the total error rate and increase the model's performance, ensemble learners typically have greater accuracies. The idea underlying ensemble modeling is the same as the one we are accustomed to using in our daily lives, such as consulting with several experts before choosing to reduce the likelihood of making a mistake or experiencing an unfavorable result.

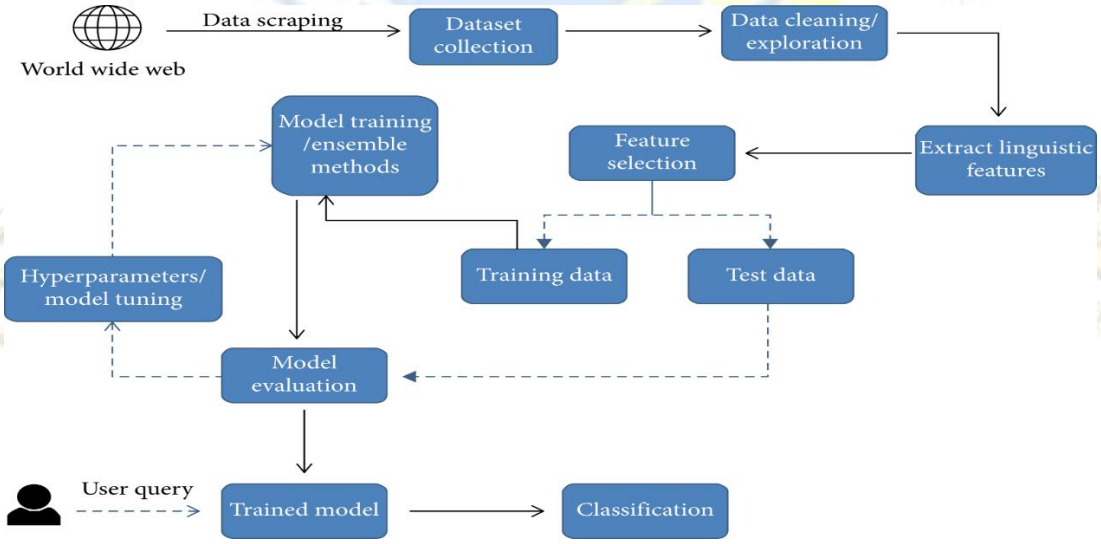
Bagging Group Classifiers:

An early ensemble technique called bootstrap aggregating, also known as bagging classifier, is primarily used to lower the variance (overfitting) over a training set. One of the most popular variations of the bagging classifier is the random forest model. To minimize overall variance in a classification issue, the bagging model intuitively chooses the class based on significant votes indicated by the number of trees, and randomly selects each tree's data using replacement sampling from the whole dataset. However, the bagging model averages across numerous estimates for regression issues.

III. OUR ASSISTS

Both supervised and unsupervised learning techniques have been used for text classification in the present fake news corpus on several occasions [20, 21]. Most of the research, however, focuses on certain datasets or domains, with the politics domain being the most well-known [10, 19, 21]. As a result, when exposed to articles from different domains, the algorithm that was trained performs less well than it should on the sort of content that it was designed for. The linguistic structure of articles varies across different domains, making it challenging to develop a general algorithm that performs well across all news domains. In this research, we suggest a machine learning ensemble strategy as a solution to the false news identification problem. Our research investigates many textual characteristics that may be utilized to differentiate authentic information from counterfeit. We use such qualities to train a variety of ensemble methods—some of which are not well-examined in the existing literature—using a mix of different machine-learning techniques. Due to the learning models' propensity to lower error rates using strategies like bagging and boosting, ensemble learners have shown to be beneficial in a wide range of applications [22]. These methods enable the

efficient and successful training of various machine learning algorithms. We also carried out in-depth tests on four real-world datasets that are accessible to the public. Using the four widely used performance indicators, the findings confirm that our suggested approach performs better than before.



IV. CONCLUSIONS

Manual news classification necessitates in-depth topic knowledge and proficiency in spotting textual irregularities. In this study, we addressed the issue of employing ensemble methods and machine learning models to classify bogus news stories. Rather than categorizing news expressly as political news, the data we utilized in our study is gathered from the World Wide Web and includes news pieces from many domains to cover most of the news. Finding textual patterns that distinguish authentic news from false pieces is the main goal of the study. Using a LIWC tool, we collected various textual characteristics from the articles and fed the feature set into the models. To get the best accuracy, the learning models underwent parameter tuning and training.

V. REFERENCES

[1] A. Douglas, "News consumption and the new electronic media," The International Journal of Press/Politics, vol. 11, no. 1, pp. 29–52, 2006.

View at: Publisher Site | Google Scholar

[2] J. Wong, "Almost all the traffic to fake news sites is from Facebook, new data show," 2016.

View at: Google Scholar

[3] D. M. J. Lazer, M. A. Baum, Y. Banker et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

View at: Publisher Site | Google Scholar

[4] S. A. García, G. G. García, M. S. Prieto, A. J. M. Guerrero, and C. R. Jiménez, "The impact of term fake news on the scientific community scientific performance and mapping in web of science," Social Sciences, vol. 9, no. 5, 2020.

View at: Google Scholar